

## Letter to the Editor

### Intragenomic Base Content Variation Is a Potential Source of Biases When Searching for Horizontally Transferred Genes

Stéphane Guindon\* and Guy Perrière†

\*Laboratoire d'Informatique de Robotique et de Microelectronique de Montpellier, Montpellier, France; and †Laboratoire de Biométrie et Biologie Évolutive–UMR Centre National de la Recherche Scientifique 5558, Université Claude Bernard–Lyon 1, Villeurbanne, France

Horizontal transfers in bacteria have been extensively studied, and most of the methods developed to identify transferred sequences are based on the assumption that exogenous genes have characteristics that differ from those shared by endogenous ones. Up to now, the most common way to look at the peculiar characteristics exhibited by “alien” sequences has been to analyze biases in synonymous codon usage (Médigue et al. 1991; Lawrence and Ochman 1997, 1998; Aravind et al. 1998; Karlin, Campbell, and Mrazek 1998; Karlin, Mrazek, and Campbell 1998; Nelson et al. 1999). The main underlying hypothesis of this approach is that synonymous codons presented by genes coming from distantly related species are different from those shared by endogenous sequences. In this paper, we show that the results obtained may be biased due to intragenomic base content variations that may occur in bacterial genomes.

As the most complete studies on horizontal transfer have been realized on *Escherichia coli* (Lawrence and Ochman 1997, 1998; Karlin, Mrazek, and Campbell 1998), we decided to center our analysis on this species. Moreover, as a detailed list of putatively transferred genes was available (Lawrence and Ochman 1998), it was possible to use it as a reference. In the study published by Lawrence and Ochman (1998), two codon usage indices were used to identify transferred genes: the codon adaptation index (CAI) (Sharp and Li 1987) and a  $\chi^2$  of codon usage (Lawrence and Ochman 1997). The selection of sequences that showed atypical values for combinations of these two indices led them to identify 755 putatively horizontally transferred genes in *E. coli* (i.e., 17.6% of the total number of genes). From among these genes, we used in our study only those that were annotated in the complete genome. This corresponds to a total of 317 genes out of 4,254 *E. coli* genes of length >150 nt. These genes are referred to as HT\* in this paper, while the others are referred to as non-HT\*. All coding sequences were extracted from the EMGLib database (Perrière, Labedan, and Bessières 2000).

To discriminate HT\* from non-HT\* genes, we used three different codon usage indices: the CAI, bias in codons (BC); (Karlin, Campbell, and Mrazek 1998), and G+C contrast in codon third positions (G+C3c). This

last index is defined as the difference between the G+C3% of a gene and the average of G+C3% calculated from the whole *E. coli* coding sequences. On the other hand, CAI and BC require a reference set of genes to be computed, and the values for the individual genes are a measure of the relatedness in codon composition to the set used. For our study, we built two different reference sets. The first set contained 17 highly expressed protein genes >150 nt: *eno*, *cspA*, *gapA*, *lpp*, *mopA*, *ompA*, *ompC*, *rplA*, *rplI*, *rplL*, *rplO*, *rpmA*, *rpmG*, *rpsA*, *rpsI*, *tufA*, and *tufB*. These genes were identified on the basis of a correspondence analysis (CA) performed on absolute codon frequencies of all *E. coli* coding sequences. To establish this set, we selected the genes with the higher factor scores on the first two axes of CA until we had obtained a pool of approximately 5,000 codons. This first set assessed biases in codon usage related to expressivity, and the two indices for it will be designated CAI and BC. The second set contained all the *E. coli* protein genes >150 nt available in the complete genome. This set assessed an “average,” or “standard,” codon usage (i.e., an absence of bias in codon composition, this relative to the complete genome of *E. coli*). The two indices for it will be designated CAItot and BCtot.

Under the assumption that horizontally transferred genes came from genomes using synonymous codons different from those exhibited in endogenous *E. coli* genes, we may expect them to present atypical values for our different indices. An appropriate way to test the significance of the difference between HT\* and non-HT\* genes is to use the nonparametric test of rank variances. This test allows the comparison of the distributions of the valuated realizations of two variables (here, HT\* and non-HT\* genes). Table 1 presents the results on HT\* gene characterization using our different indices (CAI, CAItot, BC, BCtot, and |G+C3c|). As expected, HT\* genes are associated with CAI, CAItot, BC, BCtot, and |G+C3c| values that are significantly lower than those corresponding to the other *E. coli* genes. This means that a standard HT\* sequence does not present synonymous codons used by highly expressed genes or by a standard gene from the whole genome. The rank variance tests were also highly significant for all of these indices. Among the five indices, it appears that the absolute value of G+C3c is the best measure to characterize HT\* genes compared with the remaining *E. coli* genes. Indeed, the rank variance test had the greatest significance when compared with the results obtained with the other four indices.

The distribution of the G+C3c values shown by HT\* and non-HT\* genes is given in figure 1. We can

Key words: horizontal gene transfer, codon usage, G+C% variation, DNA repair mechanism.

Address for correspondence and reprints: Guy Perrière, Laboratoire de Biométrie et Biologie Évolutive–UMR Centre National de la Recherche Scientifique 5558, Université Claude Bernard–Lyon 1, 43, Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France. E-mail: perriere@biomserv.univ-lyon1.fr.

*Mol. Biol. Evol.* 18(9):1838–1840. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

**Table 1**  
**HT\* Gene Characterization with Different Indices**  
**Measuring Biases in Codon Usage**

| Index                    | HT*   | non-HT* | MW         | RV     |
|--------------------------|-------|---------|------------|--------|
| CAI.....                 | 0.216 | 0.278   | $<10^{-4}$ | 9.145  |
| CAI <sub>tot</sub> ..... | 0.659 | 0.712   | $<10^{-4}$ | 10.555 |
| BC.....                  | 0.876 | 0.759   | $<10^{-4}$ | 8.150  |
| BC <sub>tot</sub> .....  | 0.467 | 0.350   | $<10^{-4}$ | 5.567  |
| G + C3c ....             | 0.152 | 0.056   | $<10^{-4}$ | 18.758 |

NOTE.—The mean values taken by the five indices are shown in the second column for HT\* genes and in the third column for non-HT\* genes. MW =  $P$  values from the Mann-Whitney nonparametric test of differences between means. RV = rank variance centered-reduced values (all significant at  $P = 0.05$ ).

see that it is approximately bimodal, with two peaks centered on  $-0.17$  and  $0.17$  for HT\* genes, while it is unimodal for non-HT\* genes. At first sight, this bimodal distribution corroborates the hypothesis that exogenous sequences present atypical base contents, but the imbalance largely in favor of negative values compared with positive ones is totally unexpected. Why would horizontally transferred sequences come mainly from genomes with a G+C3% lower than that of *E. coli*? Indeed, this species presents a G+C% close to 50%, and values for this parameter are approximately uniformly distributed between 25% and 75% (Sueoka 1962) in prokaryotes. To test an alternative hypothesis to horizontal transfer for explaining these low G+C3c values, we plotted the cumulative sum of the G+C3c values for all genes along the *E. coli* genome, starting from the replication origin (fig. 2). The plot shows that genes situated near the replication terminus have negative G+C3c values; in a region of 800 kb around the replication terminus, G+C3c values are highly significantly lower than those of the genes located outside this area. We performed the same analysis on eight other proteobacteria for which complete genomes were available in EM-

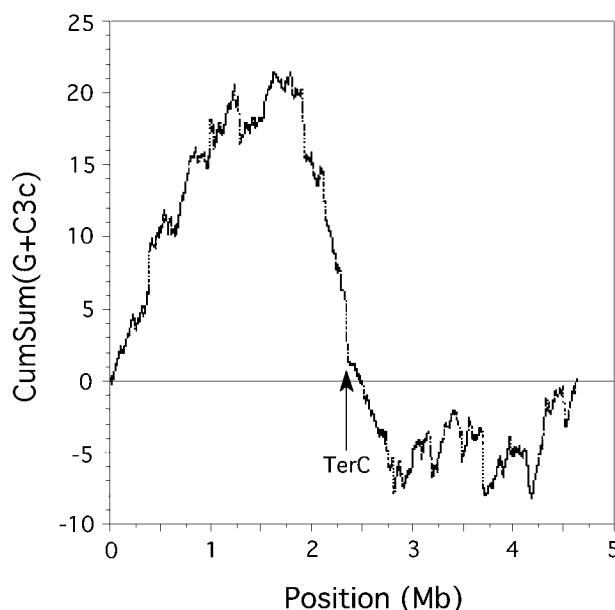


FIG. 2.—G+C3c variation along the *Escherichia coli* genome. The curve was obtained by plotting the cumulative sum of G+C3c values for all *E. coli* genes  $>150$  nt depending on their positions on the chromosome. The zero on the horizontal axis corresponds to the replication origin. A positive slope indicates a tendency to have positive G+C3c values, while a negative slope indicates a tendency to have negative G+C3c values. The genes associated with negative G+C3c values are situated near the replication terminus (TerC). The 800-kb region flanking this region contains genes presenting significantly lower G+C3c values ( $P < 10^{-4}$ ,  $t$ -test) than those in the remainder of the chromosome.

GLib: *Buchnera* sp. APS, *Campylobacter jejunii*, *Haemophilus influenzae*, *Helicobacter pylori* J99, *Neisseria meningitidis* Z2491, *Pseudomonas aeruginosa*, *Rickettsia prowazekii*, and *Xylella fastidiosa*. We used the EM-

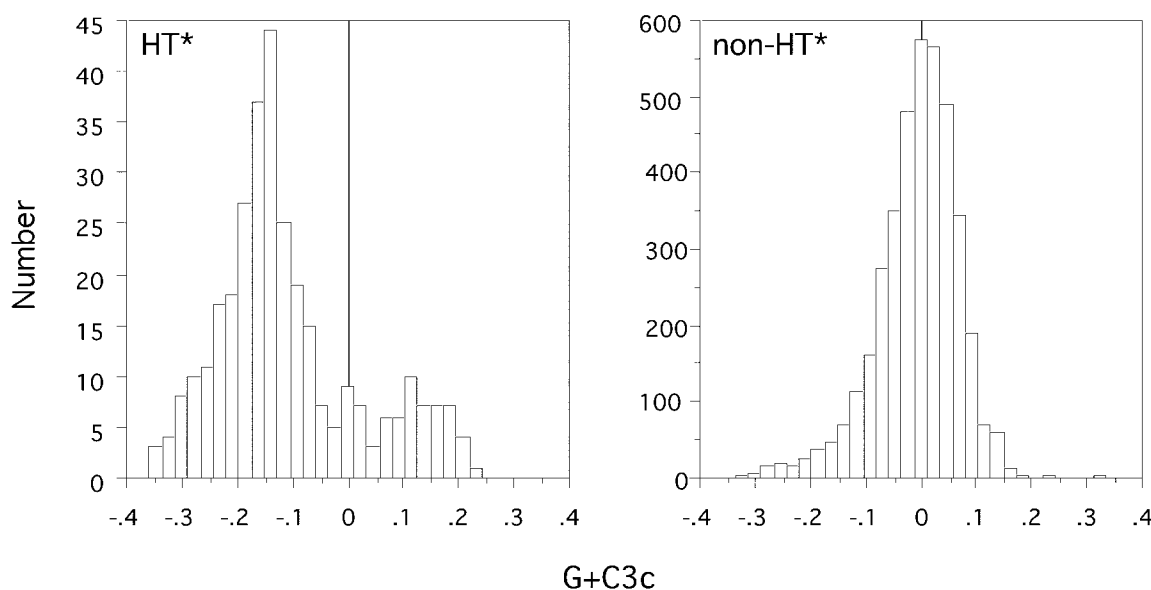


FIG. 1.—Distribution of G+C3c values for HT\* and non-HT\* genes longer than 150 nt. Among the 317 HT\* genes analyzed here, 254 (80.1%) are associated with negative G+C3c values. This proportion differs significantly from the proportion of genes with positive G+C3c values.

quences in which the replication terminus is identified through the method of Frank and Lobry (2000). Among these nine proteobacteria, five showed a clear trend of negative G+C3c values near the replication terminus: *C. jejunii*, *E. coli*, *H. pylori* J99, *R. prowazekii*, and *X. fastidiosa* (data not shown). Hence, we conclude that this trend could be common among bacteria.

As we have already pointed out, synonymous codon usage analysis has frequently been used to detect horizontally transferred genes in prokaryotic genomes. Indeed, as codon usage variation between prokaryote genomes is important (Sharp and Matassi 1994), exogenous sequences coming from a distantly related species will present a pattern different from the one shared by endogenous sequences. The problem is that such methods rely on the hypothesis that the intragenomic variation of codon usage not caused by horizontal transfers is absent or sufficiently low to be overlooked. This assumption seems unlikely, as there is a global tendency for *E. coli* genes to exhibit a decrease in G+C3% when they are close to the replication terminus. Deschavanne and Filipinski (1995) have already shown this, and these authors proposed that this variation was related to differences in DNA repair modes along the genome. Sequences close to the replication terminus are in single copies for a longer part of the cell cycle than the origin-linked genes, so they have fewer opportunities to engage in repair via homologous recombination. Due to that fact, they may resort more often to translesion synthesis, a mechanism involving the misincorporation of adenine in complement of modified nucleotides.

We suggest that the mutation drift toward A+T induced by this reparation mechanism introduces atypical biases in codon composition and that such biases will be strongest at the level of the replication terminus. Due to that fact, there is probably an overestimation of the proportion of exogenous sequences in *E. coli*, particularly near this region of the chromosome, when using methods based only on codon composition. Lawrence and Ochman (1997) already pointed out that the region corresponding to the replication terminus of *E. coli* showed the largest amount of putatively horizontally transferred genes. Indeed, 36% of HT\* genes are located within a region surrounding TerC that corresponds to only 25% of chromosome length (minutes 23–47 on the *E. coli* genetic map). Lawrence and Ochman (1997) thought that this observation reflected the high recombination levels in that region, but this deduction must be taken with caution, as Sharp (1991) showed that silent substitution rates vary significantly with chromosomal location, with genes near the replication terminus having stronger divergence. Therefore, the more distant from the replication origin the sequences are, the less they will share sequence similarity with orthologous genes. Hence, the probability of encountering minimum efficient processing segments, which are essential to recombination (Shen and Huang 1989), is lower in the terminus of the replication region. Moreover, El Karoui et al. (1999) showed that the number of Chi sites along the *E. coli* chromosome presents a depletion near TerC, and these sequences are known to stimulate DNA repair

by homologous recombination. Therefore, the fact that the replication terminus displays the most variation in chromosome size and organization could be a consequence of the lack of recombination between homologous sequences. If this hypothesis is confirmed, this kind of recombination mechanism could then be seen as a regulator of intragenomic plasticity, with high rates being associated with high conservation of genome organization.

#### LITERATURE CITED

- ARAVIND, L., R. L. TATUSOV, Y. I. WOLF, D. R. WALKER, and E. V. KOONIN. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**:442–444.
- DESCHAVANNE, P., and J. FILIPSKI. 1995. Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. *Nucleic Acids Res.* **23**:1350–1353.
- EL KAROUI, M., V. BIAUDET, S. SCHBATH, and A. GRUSS. 1999. Characteristics of Chi distribution on different bacterial genomes. *Res. Microbiol.* **150**:579–587.
- FRANK, C., and J. R. LOBRY. 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics* **16**:560–561.
- KARLIN, S., A. M. CAMPBELL, and J. MRAZEK. 1998. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32**:182–225.
- KARLIN, S., J. MRAZEK, and A. M. CAMPBELL. 1998. Codon usage in different classes of the *Escherichia coli* genome. *Mol. Microbiol.* **29**:1341–1355.
- LAWRENCE, J. G., and H. OCHMAN. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**:383–397.
- . 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95**:9413–9417.
- MÉDIGUE, C., T. ROUXEL, P. VIGIER, A. HÉNAUT, and A. DANCHIN. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**:851–856.
- NELSON, K. E., R. A. CLAYTON, S. R. GILL et al. (25 co-authors). 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**:323–329.
- PERRIÈRE, G., B. LABEDAN, and P. BESSIÈRES. 2000. EMGLib: the Enhanced Microbial Genomes Library (update 2000). *Nucleic Acids Res.* **28**:68–71.
- SHARP, P. M. 1991. Determination of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *J. Mol. Evol.* **33**:23–33.
- SHARP, P. M., and W.-H. LI. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**:1281–1295.
- SHARP, P. M., and G. MATASSI. 1994. Codon usage and genome evolution. *Curr. Opin. Genet.* **6**:851–860.
- SHEN, P., and H. V. HUANG. 1989. Effect of base pair mismatches on recombination via the RecBCD pathway. *Mol. Gen. Genet.* **218**:358–360.
- SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. *Genetics* **48**:582–592.
- WILLIAM MARTIN, reviewing editor

Accepted June 4, 2001